Systemic Data Infrastructure for Innovation Policy

Diana Hicks

Abstract— **Progress on the vision laid out in the Science of Science Policy Roadmap requires a move to system level thinking and analysis in the study of technology development. System level analysis will require systemic data infrastructure. The need for such an infrastructure is increasingly explicitly recognized at the national level. This paper will review infrastructure efforts including previous US-based infrastructure, national research documentation systems used in systemic evaluations, the Community Innovation Survey in Europe, Lattes in Brazil, the NRC ranking of US graduate programs. The strengths and weaknesses of each approach will be compared, and key issues will be identified.**

Index Terms—

I. Introduction

Progress on the vision laid out in the Science of Science Policy Roadmap requires a move to system level thinking and analysis in the study of technology development. This contribution builds on the insight that this move should be discussed explicitly as it will be a change in practice for the academic community and building a system level infrastructure will encounter foreseeable problems.

As the Roadmap details, agencies have in the past approached outcome measurement individually so a variety of approaches obtain. The interagency task group hopes to move agencies toward a more coordinated approach. A national data infrastructure that could handle the heterogeneity between fields and technologies would be a powerful support for agency coordination. In addition, it would allow agency programs to be comparatively assessed within the broader national context. Therefore, it is worthwhile to think through the practical basis of systemic infrastructure in order to foresee the challenges that lie ahead.

The sophisticated econometric models used in economic policy making are based on an expensive, comprehensive data infrastructure built and maintained over many decades by the Federal government. Using these databases, economists became used to systemic, i.e. macro, thinking. They also became used to acknowledging that although systemic data are not perfect (for example the informal economy and housework are absent) they are useful. Lacking this kind of "Cadillac" data infrastructure, the study of innovation outcomes has involved laborious construction of highly prized, career making datasets, disparaging of systemic resources as not nearly as perfect as bespoke data and case study theorizing. As close reading of the recent RAND review of outcome studies reveals, even the largest of analyses of innovation and impact have been case studies of at most a few fields or inventions [1].

The infrastructure advocated here would:

- provide a foundation for almost any quantitative analytical method: modeling, visualization, network analysis, mapping and indicator production.
- support both advancing the understanding of the research and innovation ecosystem and analysis of agency programs.
- provide time series rather than snapshots
- include evidence of linkages that reflect some part of the cumulative, networked enterprise that creates advances in knowledge and technology.
- be accessible to all interested parties, ideally free at the point of use to encourage a large and diverse community of practice.

Although the U.S. used to lead in this area, recent history suggests that other countries have surpassed the US in developing and using systemic infrastructures in S&T policy.[1] This paper will review such efforts. The paper describes an infrastructure constructed in the U.S. for tracing research funding through scientific papers to patents across a decade in time and all areas of science and technology. Also examined are national research documentation systems used in systemic evaluations, a structured CV data system, the Community Innovation Survey in Europe and the NRC national dataset on graduate programs. The paper begins by positing a vision of an ideal infrastructure and then assesses these examples of infrastructure against the ideal.

II. The Ideal Infrastructure – A Vision

In an ideal world we would be able to seamlessly connect disparate data sources to enable knowledgeable users to assess the relative strength and impact of an institution or agency

Manuscript received September 25, 2011.

D. Hicks is Department Chair of the School of Public Policy at the Georgia Institute of Technology, Atlanta, GA 30332-0345 (e-mail: dhicks@gatech.edu)

[1] It is also worth noting that the proposed Japanese funding initiative for science, technology and innovation policy includes plans to build a data infrastructure.

portfolio over time and within the context of the nation and world. This resource would also serve scholars, enabling them to build and test multivariate models, advance understanding of networks and develop visualization tools to aid comprehension of the ecosystem. Because there is heterogeneity between fields and over time, and users want to compare across fields and time, normalized metrics should be built into the system. Because government interest spans all fields of science and technology, coverage should be comprehensive. We do not want to introduce error, so the infrastructure should be built carefully, by skilled experts.

The requirements for this to happen are many:

1) Researchers have to produce output, i.e. some sort of text. Ideally, their texts contain references that link their work to predecessors. Somebody, not us, has to index that output for some other purpose.
2) We need access to the complete index, and permission to put it into our infrastructure and make it accessible for all users who want to undertake systemic analysis. Not a problem with public sources like US patents, but a chronic problem with the core of the scientific literature whose indices are owned by Thomson-Reuters and Elsevier. Though frankly, the Department of Commerce's NTIS database does not seem that open either.
3) We need to identify the people and institutions associated with each record – a daunting and expensive task whose complexity is easily underestimated.
4) The databases need to be linked through thesauri of people's names and institutional names.
5) Processes 4 and 5 must be ongoing so that the infrastructure is always current.

Points 3 through 5 highlight the need for a curator. This crucial work is always undervalued, which makes it challenging to obtain long term resources and set up the institution required.

In this ideal world we would link up texts that reported research output and texts that signified the influence of research on society and technology.

Texts related to research:

- Agency records on each grant awarded
- Papers published and their references
- Patents applied for and issued and their references in text and front page
- Technical reports and their references
- (Excluded – working papers, blogs and other forms lacking independent review before publication)[2]

Texts related to influence and outcome:

- Patent references (again)
- Policy white papers and their references
- The press, extracting any references made to scholars and researchers
- New regulations and the references made in those texts
- Congressional testimony of researchers
- Evidence based reviews of best medical practice
- New product introductions
- Download counts
- Social media mentions (Blogs, Facebook, Twitter etc.) to gauge impact on fast moving social discourse
- Etc.

With the vision in place, we can assess a selection of real infrastructures against this ideal. We can also explore the challenges inherent in mounting such an infrastructure.

III. Patent-to-Paper Citation Database

Throughout the 1980's and 1990's agencies had access to an infrastructure that did go some way towards addressing outcomes. This infrastructure was built by CHI Research, a research consulting company run by Dr. Francis Narin. CHI produced the bibliometrics for NSF's *Science & Engineering Indicators*, and as part of that contract, Narin built an analytically useful database infrastructure that allowed tracing from funding acknowledgments in papers, through to the organizations whose technology built on the funded research, i.e. whose patents referenced funded papers. The infrastructure was comprehensive across the US patent database and the *Science Citation Index*. Therefore, analyses were normalized for differences in citation rates across fields and over time. Analyses were also comparative, and so could establish how well an agency or institution was doing in comparison with others.

The steps involved in building and maintaining this infrastructure were these:

- Obtain access to the base databases, USPTO and SCI. This was unproblematic in the case of the public USPTO database, but difficult, expensive and contentious in relation to Thomson-Reuter's SCI. The base USPTO data is now even easier to obtain as it can be downloaded from data.gov. Thomson-Reuters is slightly more forthcoming since the advent of competition in the form of Scopus.
- Clean up the institutional affiliations so that the corporate owner of each patent and the institution(s) producing each paper were identified. The challenges here are keeping up with changing corporate affiliations as companies merge or fail and working at the departmental level in universities.
- Code patents and papers by geography so that analysis by Congressional district could be produced.
- Manually examine each non-patent reference in US patents, identify those that were to journal articles, standardize the reference and match to the SCI. The references also could be matched to other paper databases.
- Obtain funding acknowledgments. At the time, this was done by sending students to the library. Web of Science now includes this information in its records.

The base databases add records constantly, so if the infrastructure is to remain current, this work must be continuous. Every week CHI staff standardized between six

[2] The mores of science are unlikely to change, so to be taken seriously the infrastructure should focus on texts that have undergone some kind of review or that demonstrate the interest of people beyond the original research group.

and ten thousand non-patent references (NPRs) for US and EP patents, extracting the year, journal, author, and page from the free form text of the front-page non-patent references. The US linkage backfile covered patents from 1983 onwards and in 2001 contained 667,000+ patents with 3.78 million non-patent references, of which two million were references to scientific journal articles.

Agencies who requested studies of their impact paid for the time and effort needed to compile their data, produce an analysis and write a report. The database was built and maintained on the NSF S&EI contract and with overhead. Basic reports could be produced in one to two months. Francis Narin has since retired, and with his departure the unique vision of the value of this kind of infrastructure has faded. This infrastructure is high cost but of limited value to private sector clients, therefore it no longer exists.

There have been other efforts to mount a similar infrastructure. DOE attempted to collect and link the patents and paper abstracts of its PI's, but this project ended after a departmental reorganization. Academics have been somewhat more successful. NBER has produced two versions of a cleaned patent database although this has never been as current, accurate in institutional affiliation or inclusive of paper-linkages as the CHI databases. The current Fleming data contains the patent-paper linkage element, but is not comprehensive and was built on an NSF grant, which suggests it will disappear after a short time. NSF has also funded a database linking effort at UCLA the first fruits of which are available to NBER members. The NIH has been most successful. Its $100 million per year expenditure on the PubMed infrastructure provides a base to link to NIH grants and NIH related patents in various systems including RePORTER, SPIRES and e-SPA. Interest is obviously building in unified database solutions to analysis of the science system. However, each solution has particular limitations and none currently approaches the ideal of universally accessible, comprehensive coverage of the US science ecosystem. None of this should come as a surprise because key data elements are privately owned and because database curation is expensive and does not fit well into the incentive system faced by professors or into agency definitions of "research activity"[2].

To compare this genre to the ideal, we examine the strongest of these efforts, the CHI databases. These fell short of the ideal along several dimensions. Because the link to funding was made from acknowledgements on papers, the data were incomplete and were at the agency level. Thus analyses could be used for advocacy, where assessing the strength of the agency as a whole is helpful but not for program management which requires program level data. The problem here would be obtaining government-wide data on grants (the experience of Radius being relevant). Once an open, accurate, functional database of government R&D grants and contracts was built, individual people would need to be uniquely identified and matched to authors and inventors. For completeness, the NTIS database of technical reports would need to be opened up, and the same cleaning of meta-data conducted.

On the outcome side, the situation is even worse. CHI used patents at the firm level which is convincing because one can say that the firm's products are built on the firm's technology, which is more or less documented in its patents, and patents document some of the links to research in their references to papers. In an ideal world we would have documentation of links between research and individual products. This only exists in pharmaceuticals because FDA regulations require that product applications be supported by extensive textual evidence with a scientific aura – i.e. references.

To truly embrace the full spectrum of outcomes, we would need to incorporate many more types of text: regulations, best practice recommendations, press, Congressional testimony, policy white papers, mentions in social media etc. All of these would need to be scraped for mentions of researchers which would be catalogued and linked, again by researcher name, to the database of research grants and outputs.

IV. National Research Documentation Systems

The difficulty in constructing an evaluation infrastructure is that Thomson-Reuters or Elsevier own the comprehensive databases of journal articles and citations.[3] To build an evaluation infrastructure, one needs access to all the records in a database, and then one needs to do cleaning and indicator construction. This is not a model private database providers support. As a result, CHI faced chronic difficulties in constructing its infrastructure. In the medical area, NIH spends a great deal of money to build and maintain PubMed which as a public database is free at the point of use. Fleming's data uses PubMed. There is no equivalent for the rest of science, the social sciences or the humanities. Ironically, countries that have felt excluded from Web of Science have moved ahead with their own electronic publication databases using open access and so may be better able to build evaluation infrastructures on top of these databases. In Latin America there is SciELO, Scientific Electronic Library Online, a federation of electronic journal infrastructures that meet a centrally defined standard of excellence in journal publishing (scielo.org). SciELO's site not only provides access to 250,000 articles from 660 journals, but also offers basic bibliometric statistics. Similarly, in Africa there is African Journals Online (ajol.info) hosting 46,000 articles from 396 peer reviewed journals. China has built the CSCI, and CSSCI (Chinese versions of the SCI and SSCI), administered by CAS National Science Library and Nanjing University respectively.

An alternate route to a similar end is a national research documentation system. These have been created in several smaller countries – Norway, Australia and most recently Denmark. These infrastructures have emerged to support national, metrics-based evaluation of university research. In these countries it is possible for the national government to mandate university participation as a condition of receiving their block grant. In the U.S. this is not possible. However,

[3] This discussion is simplified for the sake of brevity. A complete discussion would probe the strengths and weaknesses of other indexes such as Inspec, ChemAbstracts, Google Scholar, Xarchiv, CiteSeer, etc.

the model may have relevance for agency intra-mural institutions.

In national research documentation systems, universities submit bibliographic records of their publications and are responsible for data quality. In the Norwegian system, the agency validates and standardizes bibliographic records submitted by universities. This involves creating and updating an authority file of allowed publication channels. Data from Thomson Reuters and the Norwegian national library are imported to verify and standardize records. The authority file standardizes names of publication channels, document types, and institutional affiliations of authors. The work by the agency recognizes and addresses known accuracy problem in submitted data. Publications are differentiated according to a 2-4 level classification of the quality of the publication venue. Weighted publication counts or publication distributions across the levels are then produced. The system is somewhat costly. Full cost includes that born by universities in submission and by the agency in validation.

The base model contains no impact measures – i.e. citations - but the Australian version extends the system. Australia will buy data from Scopus, and for papers published in journals indexed in Scopus, citation counts will be produced. This will serve as an additional dimension in the evaluation in addition to the distribution of papers across journal level.

A key limitation of this model is that it does not serve to capture the connection to grants or outcomes - economic, social, health and environmental benefits or technological developments. These governments first sought to evaluate and increase scientific quality, and the infrastructure serves that purpose. One could imagine an analogous system that might go some way to assessing outcomes. For example, intra mural laboratories could be required to submit not only information on journal articles, but also patent information, information on impact on the "enlightenment" literature, i.e publications that reach industry and decision makers (*Wall Street Journal, New York Times*, trade press etc.) and mentions in social media. Or end of grant reporting could be changed from narrative texts to entry of structured data on outputs, enlightenment literature, patents, people employed etc. (It would probably not be ideal to require structured data entry in addition to traditional narrative reports.)

V. LATTES

One way to move toward this goal is to work with CVs. Brazil has built an open access web-based CV infrastructure that serves as a model for investigator level data collection. The Brazilian Lattes website provides the tools for researchers and students to register and build or update a CV at any time[4]. The system structures the entry process to produce usable information. The Lattes CV system was developed for the National Council for Scientific and Technological Development (CNPq), and is used by the MCT [Science and Technology Ministry], FINEP [Projects and Studies Financing], CAPES/MEC [Personal Improvement Coordination/Ministry of Education], and the Brazilian scientific community. In the CNPq, the information is used to evaluate candidates for scholarships and/or research support; select consultants, members of committees and advising groups; and evaluate Brazilian post-graduate education and research programs. Universities are allowed access to the data for their own indicators and analysis. Information scientists develop tools to mine the data and a small amount of bibliometric work has been done using the data. Researchers can use the Lattes CV as their website (though if Lattes mounted working papers for download it would be more useful for researchers).

As the system matured, it has developed. First, Lattes CVs became mandatory in 2002 for all researchers and students that deal with the Council. Absence of a Lattes CV causes impediments to payments and grant renewals. More recently, CNPq has improved data validation. CNPq has bought Thomson Reuters data so that CVs have online access to the number of citations in Web of Science for the articles registered in Lattes with their digital document identifiers (DOI). Researchers can also see who has cited them. To retrieve the number of citations, researchers must ensure that the DOI, the journal, volume, issue and pages (first and last) are registered properly. CNPq also uses data from Federal Revenue of Brazil to validate personal data such as the Individual Taxpayer Registry (CPF) document, name, date or birth and affiliations, preventing the creation of fake new CVs. This validation is new and is not applied retroactively to CVs already in existence.

The strength of the Lattes system is its focus on the individual. S&T data infrastructures have largely functioned at the organisational level. This is true of the CHI databases, national research documentation systems and the CIS (below). Such infrastructures are useful for producing S&T indicators, which are largely comparisons of nations. Organizational level data can go some way to support understanding of the ecosystem and study of policy effects. They also serve governments deciding how to divide block grants among universities. As was mentioned above, working at the firm level helps link patents to products and economic success since firms have a patent oeuvre, a product portfolio and performance metrics. So the organizational level is extremely useful.

Today there are pressures to move to the lowest possible level, the individual investigator. These pressures arise both in the scholarly community and among government agencies. Methodological interest in agent based modelling and network analysis generate a need to create infrastructures that track the contributions of individual investigators. Federal funding agencies do not fund institutions, they fund investigators, and to understand the impact their programs are having, they need to be able to track the contributions of people, not institutions. The strength of the Lattes approach is the individual basis of the data. In other systems, accurate identification of the individual responsible for or mentioned in a text is extremely challenging. People are tricky: a lot of them have the same name; they change names; they change jobs; they misspell the names of others; they make errors when entering names in

[4] Very little information was found on the Lattes system, this section is basically taken from the Lattes website: http://lattes.cnpq.br/english/index.htm

databases; they choose different conventions for representing names when setting up different databases, etc. The SciSIP community is actively investigating solutions to these problems. Although doing the topic justice would require a review in itself, suffice it to say that some suggest scholars adopt an ID number and associate it with all their output while others work on algorithmic solutions. Requiring researchers to enter all their own data is the ultimate solution to this problem, though it brings with it another set of challenges.

The challenges of the Lattes system lie in linkages and data quality. CVs do not naturally contain much linkage information. The system could overcome some of this by requiring co-authors with Lattes CVs to be identified when a paper is entered. This provides some network information. The pure Lattes system also lacks impact indicators, but a Lattes-Web of Science/Scopus hybrid can solve this problem and this is the direction Brazil is taking the system. The Lattes system could also go some way toward incorporating societal impact measures if a section on public dissemination were built.

Data quality is also a challenge for a Lattes system. If the system is used for assessment, there will be incentive to falsify and inflate the CV which must be countered in the design and operation of the system. Validating identity against IRS records would be a gold standard. Presumably institutional affiliation could be validated through cross-checks with systems such as Fastlane. Journal article entries could be validated by cross-checking with Web of Science or Scopus data. Another method of controlling what is listed as a journal article is only allowing recognized peer reviewed journals to be included (the authority list approach used in national research documentation systems).

Strangely for a vehicle that contains as much information as does a CV, context is lacking and this compromises the utility of the data. For example, if a researcher gets a grant and this supports work leading to some papers, several conference presentations, testimony to Congress and a patent, each of these items will be on the CV, but they will be in different sections and the funding agency will not be able to determine which items on the CV are associated with their grant [3]. Contrast this with working with funding acknowledgements in papers.

Other limits of a Lattes system are the potential holes, the national basis, and the shallow content . Not every active researcher is engaged with a Federal agency, or is currently engaged. This will lead to absent or outdated material. For example, coverage of humanities and some areas of social science will be sparse. Further a system driven by national funding agencies will be domestic, eliminating the potential for international comparison inherent in comprehensive infrastructures based on comprehensive databases (also true of national research documentation systems, NRC and CIS – see below). A Lattes system is also limited by having only the title of a paper. Even having an abstract increases the sophistication of the analysis, and full text databases allow much more sophisticated analysis based on content (including references, acknowledgements etc.).

VI. NRC Exercise

In the U.S. the NRC gathered extensive data from universities via a survey in order to rank, or rather to avoid ranking, graduate programs. The NRC has built an accessible data infrastructure which includes survey responses as well as bibliometric data purchased from Thomson-Reuters. As the NRC's 1995 ranking, based on a reputational survey, was seen as too "soft", the current effort collected quantitative data. A small reputational survey was conducted, and a regression analysis used to identify a weighted mix of quantitative variables that best predicted reputational judgments. The NRC required from the 5000 participating programs at 212 universities submission of information on the 48 variables to be included in the ranking formula. The result is almost 1/4 million data points. The 48 variables concern institutional characteristics (i.e. total research expenditure, characteristics of library, childcare and health insurance availability, university housing for PhD students etc.); doctoral program characteristics (i.e. size, time to degree, financial support, facilities for PhD students, test scores, support provided, employment destinations etc.), and program faculty (size, demographics, awards, bibliometrics etc.) (NRC, 2004, Table 4.1). For the bibliometrics, the NRC compiled full bibliographies of *SCI* indexed papers and their citations from Thomson-Scientific and used this information to calculate bibliometric variables at the departmental level: publications/faculty, and citations/publication.

The NRC method was elaborate, and this had a cost. Planning for the exercise began in 2000; it was originally scheduled for 2003-2004 and slated to cost $5 million (direct cost only); it was actually conducted in 2005-2006 for release in 2007. The results were released in September 2010. Questions are being raised about the accuracy of the data and whether they are too dated to be informative.

Again, this data infrastructure does not index the origin of grants or outcomes such as economic, social, health and environmental benefits or technological developments from the work in U.S. graduate programs. Presumably the survey method could be refocused to query on these items.

VII. Community Innovation Survey

The European Community Innovation Survey or CIS provides a model of an outcome focused, survey-based data infrastructure.[5] The survey is administered to a random, stratified sample of firms. The CIS began in 1993 and has been conducted five times, the latest being CIS-5 in 2007. The survey is conducted by 30 countries, with Eurostat working to harmonize the questionnaire and methodology.

The CIS collects information about product, process, organizational and marketing innovation during the three years prior. Questions cover new or significantly improved goods or services or implementation of new or significantly improved processes, logistics or distribution methods. The survey provides a basis for statistics on the occurrence of innovation across Europe by types of firms, geographic location, type of

[5] The CIS discussion is based upon [4].

innovation (process, product, organizational etc.), innovative activities (R&D, acquisition of advanced equipment or software, training, marketing etc.), amount spent on innovation, and effects of innovation on goods or services. Questions are asked about public support and public information sources, but we learn only what types of support or institutions were helpful. To be useful for agency analysis, firms would have to be asked which agencies and which institutions were supportive.

CIS data is widely used in European studies of innovation. The survey now provides time series indicators of innovation, though panel data is not possible since different firms are surveyed in each round. Descriptive analysis of CIS data has established that innovation is spread widely in the economy, beyond the relatively few firms with substantial R&D expenditure. Formal R&D expenditure accounts for just 20% of expenditure on innovation. Diffusion of new technology, through purchase of advanced equipment for example, has been shown to be a very important route for firms to update their products and processes. Multivariate analysis has been used to examine, for example, the effects of innovation strategies on performance. The survey data have enabled quite a bit of analyses of the relationship between government support for innovation and innovation outcomes. Both the incidence of innovation with government support and the additionality of government support have been examined.

There are several obstacles to greater impact of CIS work. First, the analytical value of the data is limited because any access to micro data requires project-by-project approval of every country whose data is used. Second, academics have not seriously engaged with policy makers and their agendas. Third, constraints on length of questionnaire limit the depth of data that is collected, for example, agencies providing support are not named. Constraints to do with anonymity make it difficult to link the firm level data with other sources providing information on business performance. The lack of panel data is also a serious limitation with regard to policy relevant analysis.

In 2010 NSF released the results from the first comparable US survey – the 2008 BRDIS.[6] BRDIS is a joint effort of NSF and the U.S. Census Bureau. BRDIS has been designed to collect a wide range of data on business R&D and innovation activities in the United States, including on topics that were not addressed by its predecessor, the Survey of Industrial Research and Development. The sample of companies for BRDIS is selected to represent all for-profit companies in the United States with five or more domestic employees, publicly or privately held. The resulting sample provides statistical estimates for the population of companies that perform or fund R&D or engage in innovative activities in the United States. For the 2008 BRDIS survey, 39,553 companies were sampled, representing 1,926,012 companies in the population. When these preliminary data were tabulated, the overall response rate was 77%, and the response rate for the top 500 domestic R&D-performing companies was 93%.

[6] BRDIS discussion taken from: NSF Releases New Statistics on Business Innovation, NSF 11-300, October 2010 - http://www.nsf.gov/statistics/infbrief/nsf11300/

In comparing the CIS and BRDIS, note that in contrast with CIS emphasis, NSF states that one of the clearest findings in the BRDIS data is companies with R&D (either performing R&D or funding others to perform R&D) exhibit far higher rates of innovation than do non-R&D companies. There will be fewer obstacles to analytical use of the BRDIS than there are to use of CIS data. Presumably, BRDIS data will be housed in Census Data Centers which will allow qualified scholars access to the full data set and the opportunity to link it with other business data.

As infrastructure the CIS/BRDIS approach excels where the grant-paper-patent approach is weak – it is outcome focused, comprehensive and encompasses non-R&D based innovation including that in services, organization, marketing etc. Very valuable time series indicators can be produced, akin to traditional economic indicators. However, this approach is not the best one for enabling agencies to track the impact of individual R&D programs. The approach also may fall prey to standard weaknesses of surveys – declining response rates and doubts as to whether everybody filling in the survey can provide accurate answers to questions such as:

> *whether during the three years prior the enterprise co-operated on any innovation activity with one or more of seven types of institutions including "Government or public research institutes" in: a) their country, b) other Europe, c) U.S., d) other countries (question 6.3, CIS 4th harmonised survey questionnaire).*

VIII. Classification in S&T Infrastructures

Information infrastructures require classification systems so that individual records can be aggregated for analysis. Classification systems emphasize continuity and so have trouble keeping up with reality, especially when they are used to examine an activity such as research that seeks to invent entirely new categories. What are deemed useful categories varies with the users' perspectives, so coordinating among agencies will mean developing a consensus classification system, and the difficulty of doing this should not be underestimated. Different classification systems and issues are involved in each of the infrastructures reviewed here.

CHI Research used NSF's classification of scientific fields for papers and a custom aggregation of patent classes. Patent offices classify patents based on the "art" involved. Everyone else is interested in the industry represented by the technology. Therefore innovation analysts who works with patents must develop a thesaurus to translate from patent classes based on "art", to the industrial classification scheme - NAICS (formerly SIC). Patent classification schemes developed for analysis usually have about 30 categories and correspond roughly to industries at the NAICS 3 digit level.

The design of a research documentation system begins with a consultative design process to specify the classification systems (fields and journal quality levels) and the system's boundaries (journal list). This can be a contentious process and reaching inter-agency consensus would be difficult, but necessary if coordination and comparison were desired.

Lattes is challenged to fit everybody into a single set of boxes so that counts might be undertaken. The system includes field classifications, and at a high level researchers are presented with a list of choices. At the more specialist level, researchers can use predetermined categories or enter their own. Researcher invented categories produce a lot of scatter in field identification at the specialist level, limiting analytical potential. A similar set of issues would apply to institutional and departmental identification. If researchers can be forced to pick a department and institution from a closed list, an analytically useful dataset will be created. If freedom is allowed in entering affiliation, a large amount of data cleaning will be needed prior to any analysis.

The NRC used a 62 category field taxonomy that included established fields as well as those designated as emerging (feminist & gender studies, film, rhetoric, information science, nuclear engineering etc.). The NRC scheme has categories for French, German and Spanish and so would have no way to accommodate a department of modern languages. International affairs or the emerging area of digital media/gaming studies do not appear, even among emerging fields. Because older fields are secure in taxonomies, traditionally defined departments are advantaged in evaluation/rankings in comparison to newer, interdisciplinary areas. To the extent that university administrators and students make decisions based on the NRC rankings, newer departments will be hurt.

The CIS asked respondents to classify their firms into the European equivalent of NAICS at the 4-digit level, that is into one of 500 categories. Choosing a fine-grained category in a traditional industrial classification might present problems for very innovative firms. In the beginning new industries are too small to warrant their own category within an economy-wide classification system that has been in existence for many decades and seeks to maintain continuity with the past. For example, business model innovation is reshaping our economy at an even more fundamental level than technological innovation as goods and services are sold together, tasks become tradable in addition to products and the focus shifts to markets and consumers rather than producers. The conceptual basis of the NAICS industrial classification scheme reified the agriculture OR manufacturing OR service framework and has great difficulty adapting.

Systemic infrastructure will require hidden and underappreciated foundations - classification systems such as NAICS or NSF's scientific field classification. Entities that fall between categories will be disadvantaged. The UK RAE was criticized for penalizing interdisciplinary departments. Non-neoclassical economists have struggled in evaluation systems in which the "economics" category is defined by neoclassical economists. Firms that don't fit neatly into categories have been shown to be disadvantaged in credit scoring [5]. These problems are of the same type that led the government to allow people to check multiple race/ethnicity categories on their Census forms.

Of classification schemes, Bowker and Star commented "architecture becomes archaeology over time"[6]. As the NRC example suggested, category systems can easily exclude the most innovative work. In a fast moving economy subject to fundamental change, the study of innovation will be hampered by schemes that privilege areas prominent when they were first devised. Although revisions take place, it can be argued that they take care of certain usual suspects, semiconductors and biotechnology for example, upon which we then build our entire understanding of innovation, to the neglect of obviously innovative areas such as photonics and gaming which are hidden in current classifications. On the other hand, allowing a free-for-all data entry system results in data that cannot be analyzed. True systemic understanding will need to address this problem.

IX. Summary

A truly systemic database infrastructure built to track research and innovation and their societal outcomes would be valuable both for agencies managing programs and for academics advancing fundamental understandings of the science and technology ecosystem. Here we have imagined an ideal system, as well as noted the development of parts of that ideal system.

There is a potential alternative to the approaches here that gets very close to outcomes, and has never been tried. The RAND report on outcomes makes clear that relationships are essential for research results to transition to application. The problem is that relationships do not always leave traces in written records, so to analyze them we must find a way to surface them. Noting that the flip side of a productive relationship with industry is a conflict of interest, Toby Smith of the AAU points out that mandatory conflict of interest reporting, implemented through a standardized, structured data entry interface could create traces of outcome related relationships that could be analyzed. There are privacy concerns, so perhaps Census Data Centers would need to house the resulting data and manage its analysis.

The examples reviewed here exhibit a full range of funding and institutional mechanisms, including grants (Fleming), long term intramural institutional funding (PubMed), agency budgets (national research documentation systems, Lattes) and were produced by government agencies, libraries, universities, non-profits and firms. An ideal system would combine the long term commitment and level of resources marshaled by NIH's PubMed, with a mandate to encompass all of science and technology (like Web of Science, Scopus or Lattes), in an independent non-profit focused on building and curating an open infrastructure for the use of scholars and government agencies (somewhat like ICPSR, though that is housed in a university). Such an entity could pay high enough salaries to hire the most talented programmers as well as employing data cleaners and curators in tasks that bring no academic glory.

There are of course substantial challenges in building any of these infrastructures – data owners refusing access, researchers resisting mandatory data entry requirements, lack of a social science research funder able to commit substantial resources guaranteed over the long term, lack of interest in the social science research community accustomed to case study level work. Any method chosen will be imperfect, and this will be

obvious to traditionalists seeking to block an initiative. Nevertheless, there are encouraging shifts. Large scale databases are becoming more accepted in social sciences, and cheaper to mount and manage. The popularity of network analyses leads naturally to consideration of entire systems. And interest among agencies advances the conversation.

REFERENCES

[1] S. Marjanovic, S. Hanney and S. Wooding. "A historical reflection on research evaluation studies, their recurrent themes and challenges." RAND Corporation, Santa Monica, CA: 2009.

[2] P. N. Schofield, J. Eppig, E. Huala, M. Hrabe de Angelis, M. Harvey, D. Davidson, T. Weaver, S. Brown, D. Smedley, N. Rosenthal, K. Schughart, V. Aidinis, G. Tocchini-Valentini and J. M. Hancock, "Sustaining the Data and Bioresource Commons," *Science*, vol. 330, pp. 592-593, Oct. 2010.

[3] F. M. Silva and J. W. Smit, "Information organization in open electronic systems of Scientific and Technological Information: analysis of the Lattes Database," *Perspect. ciênc. inf.* vol.14 no.1 Belo Horizonte Jan./Apr., doi: 10.1590/S1413-99362009000100007

[4] A. Arundel, C. Bordoy, P. Mohnen and K. Smith, 2008, "Innovation surveys and policy: lessons from the CIS," in *Innovation Policy in Europe: Measurement and Strategy,* C. Nauwelaers and R. Wintjes Eds. Cheltenham, UK. Edward Elgar, 2008, ch. 1, pp. 3-28.

[5] M. Ruef and K. Patterson. "Credit and classification: the impact of industry boundaries in nineteenth-century America." *Administrative Science Quarterly*, vol. 54, pp. 486–520, 2009.

[6] G. Bowker and S. L. Star, *Sorting Things Out*. Boston, MA: MIT Press, 2000.

Diana Hicks is Chair of the School of Public Policy, Georgia Institute of Technology, Atlanta GA, USA. She obtained her DPhil in Science and Technology Policy Studies from SPRU, University of Sussex, UK. For almost 10 years she was on the faculty of SPRU. Between 1998 and 2003 she was the Senior Policy Analyst at CHI Research, Inc. where she conducted numerous policy analyses for government agencies based on empirical information in patent and paper databases. Her work has been supported by and has informed policy makers in the U.S., Europe and Japan. Prof. Hicks has also taught at the Haas School of Business at the University of California, Berkeley, SPRU and worked at the National Institute of Science and Technology Policy (NISTEP) in Tokyo.